WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Fake News Detection Using Machine Learning

Neha Shah

School of Engineering, P P Savani University, kosamba, Surat, 394125, India

A B S T R A C T

The rise of fake news and misinformation has become a significant concern in today's digital age, impacting public perception and decision-making. To combat this, the project introduces a machine learning-based solution aimed at detecting fake news through an intuitive web application. The solution allows users to submit news articles, which are then classified as real or fake based on the trained machine learning models. The dataset, curated through web scraping, includes both fake and real news articles, ensuring a balanced input for training. The data goes through rigorous cleaning and preprocessing, followed by feature extraction using TF-IDF vectorizer and n-gram techniques, which help in capturing key linguistic patterns. Machine learning models such as Logistic Regression, Decision Tree Classifier, and Random Forest were implemented to build a highly accurate classification system. The web application serves as an automated platform where users can verify news credibility in real-time, providing an accessible tool to reduce the spread of misinformation. This solution offers a streamlined, user-friendly experience, helping users stay informed with authentic news content while leveraging machine learning for enhanced accuracy and efficiency.

Keywords: Fake News Detection, Machine Learning, TF-IDF, Logistic Regression, Decision Tree, Random Forest

1. Introduction

In today's digital era, the rapid dissemination of information through online platforms has amplified the spread of misinformation and fake news. This growing challenge not only distorts public perception but also poses serious social, political, and economic consequences. In response, this project introduces a Fake News Detection System, a machine learning-based web application developed to identify and flag deceptive news content.

The primary objective of the system is to offer a reliable and user-friendly platform that allows users to verify the authenticity of news articles in real time. The system employs advanced machine learning techniques—such as TF-IDF vectorization and n-gram analysis—trained on a well-curated dataset collected through web scraping. These models enable accurate classification of news as real or fake based on linguistic patterns and contextual cues.

The Fake News Detection System offers several key features that make it both effective and accessible for users. It enables accurate news verification by allowing users to input news content and instantly receive feedback on its credibility. Through a real-time, web-based interface, users are empowered to make quick, informed decisions about the authenticity of the information they encounter. The system ensures robust data processing by implementing thorough data cleaning and preprocessing techniques, which significantly enhance model accuracy and enable efficient handling of large volumes of news data. Moreover, its user-friendly design ensures that people from diverse backgrounds can easily navigate and utilize the platform. By integrating these features, the project not only addresses the pressing issue of misinformation but also contributes to the promotion of truthful content, improved digital awareness, and greater trust in the information consumed online.

2. Related Work

Numerous approaches utilizing machine learning techniques have been developed as a result of the considerable interest that the automated identification of fake news has received from a wide range of disciplines. Finding efficient detection techniques is necessary since the spread of fake news, especially on social media, has increased worries about its possibly harmful effects on society [1][2]. Content-based analysis, which examined textual characteristics such writing style, sentiment, and factual accuracy, was frequently used by early detection systems [3]. More recent research examines how fake news is dynamic as it spreads across social networks, making identification more difficult and necessitating the use of more advanced analytical methods [1]. Depending on the dataset and feature representation used, machine learning algorithms have been widely used to identify whether news stories are authentic or fake, with varying degrees of success [4]. Since they provide a flexible toolkit for differentiating genuine content from misinformation, natural language processing techniques have grown in popularity in this field [5]. Inspired by artificial neural networks and the structure of the human brain, deep learning techniques have become powerful tools that can automatically learn hierarchical data representations and extract complex patterns

that are essential for precise predictions [6]. However, the constantly changing strategies used by disinformation propagators and the inherent biases in training datasets make it difficult to develop models that are universally applicable [2].

3. Methodology

***1. Data* Collection:** The first stage involved gathering data from newswebsites. Web scraping was performed using the BeautifulSoup library tocollect news articles from various sources, ensuring a diverse dataset ofgenuine and fake news. This step aimed to build a comprehensive datasetfor model training and evaluation.

2. Data Preprocessing: After data collection, the text data underwentpreprocessing to make it suitable for analysis. This involved cleaning thetext byRemoving special characters, numbers, and punctuation tostandardize the content. Converting all text to lowercase to maintain consistency.Eliminating stopwords (commonly used words such as "the" and "is")that do not contribute to the meaning of the text.Applying techniques like tokenization to break the text intoindividual words or phrases. These preprocessing steps improvedthe quality of the input data and prepared it for feature extraction.

3. Feature Extraction: To transform the cleaned text data into numericalform, feature extraction techniques were applied: Bag of Words (BoW): Converted text into word frequency counts without considering the order of words, TF-IDF (Term Frequency-Inverse Document Frequency): Measuredthe importance of a word in a document relative to the entire dataset,highlighting words that are more informative. Count Vectorization: Counted the occurrences of words within thedocuments. After comparing the three methods, TF-IDF combinedwith N-grams was selected for this project to capture word sequencesand contextual relevance effectively.

4. Model Building: With the features extracted, machine learning modelswere trained using different algorithms:

Linear Regression: Although commonly used for regression tasks, itwas adapted here for classification by converting predictions intobinary classes.

Decision Tree Classifier: Used for its ability to create decision rulesthat classify the news based on features.

Random Forest: An ensemble learning technique that combinedmultiple decision trees to improve prediction accuracy and reduceoverfitting.

5. Model Evaluation: The trained models were evaluated based on theirperformance metrics:

Accuracy Score: Measured the proportion of correctly classifiedinstances.

Classification Report: Provided detailed insights into precision,recall, and F1-score for each class (fake or genuine). The dataset wassplit into training and testing sets to evaluate the models'generalization ability, ensuring the system performs well on unseendata.

6. Deployment: The final step was deploying the model through a webinterface built with Flask. The deployment involved:

Integrating the trained model to enable real-time fake newsdetection.Creating a user-friendly interface where users could input newsarticles or headlines and receive classification results.Ensuring the web application could handle user requests efficientlyand deliver accurate predictions.

4. Results

To evaluate the performance of the fake news detection system, three machine learning models—Logistic Regression, Decision Tree, and Random Forest—were trained and tested using a labeled dataset. Each model was assessed using key evaluation metrics including accuracy, precision, recall, and F1-score, which collectively reflect the quality and reliability of the predictions.The Logistic Regression model, while simple and efficient, achieved a respectable accuracy of 93%. It maintained balanced values across precision, recall, and F1-score, making it a suitable choice for quick, baseline predictions.The Decision Tree Classifier slightly improved upon the results of Logistic Regression, achieving 94% in all metrics. Its interpretability and decision rule-based approach made it useful for understanding the factors influencing the classification outcomes. However, the Random Forest Classifier outperformed both models, achieving the highest scores across all performance metrics. With an accuracy of 96%, and similarly high precision, recall, and F1-score values, Random Forest demonstrated its strength in capturing complex patterns in textual data. Its ensemble nature—combining multiple decision trees—helped minimize overfitting while improving generalization on unseen data.These results indicate that Random Forest is the most effective model for fake news detection in this study. It provides a strong balance between performance and robustness, making it the ideal choice for deployment in the web application for real-time fake news classification

Table 1 –Result Analysis

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	**93%**	**0.93**	**0.93**	**0.93**
Decision Tree	**94%**	**0.94**	**0.94**	**0.94**
Random Forest	**96%**	**0.96**	**0.96**	**0.96**

5. Conclusion

This research presents a comprehensive end-to-end approach for detecting fake news using machine learning techniques. Starting from data collection via web scraping, through preprocessing and feature extraction, and culminating in the training, evaluation, and deployment of classification models, the study effectively demonstrates how automation can support the battle against misinformation.Among the models implemented—Logistic Regression, Decision Tree, and Random Forest—the Random Forest Classifier emerged as the most effective, delivering the highest accuracy and balanced performance across precision, recall, and F1-score. Its ensemble learning mechanism provided robustness and adaptability in handling diverse and complex textual data, making it highly suitable for real-world applications.To make the solution accessible and actionable, a lightweight web application was developed using

Flask, offering users a simple interface to verify news authenticity in real-time. The application's responsiveness and intuitive design make it suitable for integration into media literacy initiatives, educational platforms, and browser extensions, promoting responsible information consumption.Overall, the paper successfully demonstrates the potential of machine learning in addressing one of today's critical digital challenges. By providing a practical, scalable, and efficient fake news detection system, this work contributes to strengthening information integrity and empowering users to make informed decisions in the digital information ecosystem.

References

[1]Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. Social Network Analysis and Mining, 13(1), 30.

[2]Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications, 4, 100032.

[3]Prachi, N. N., Habibullah, M., Rafi, M. E. H., Alam, E., & Khan, R. (2022). Detection of fake news using machine learning and natural language processing algorithms. Journal of Advances in Information Technology, 13(6).

[4]He, L., Hu, S., & Pei, A. (2023). Debunking Disinformation: Revolutionizing Truth with NLP in Fake News Detection. arXiv preprint arXiv:2308.16328..

[5]Katsaros, D., Stavropoulos, G., & Papakostas, D. (2019, October). Which machine learning paradigm for fake news detection?. In IEEE/WIC/ACM International Conference on Web Intelligence (pp. 383-387)..

[6]Berrondo-Otermin, M., & Sarasa-Cabezuelo, A. (2023). Application of artificial intelligence techniques to detect fake news: A review. Electronics, 12(24), 5041.